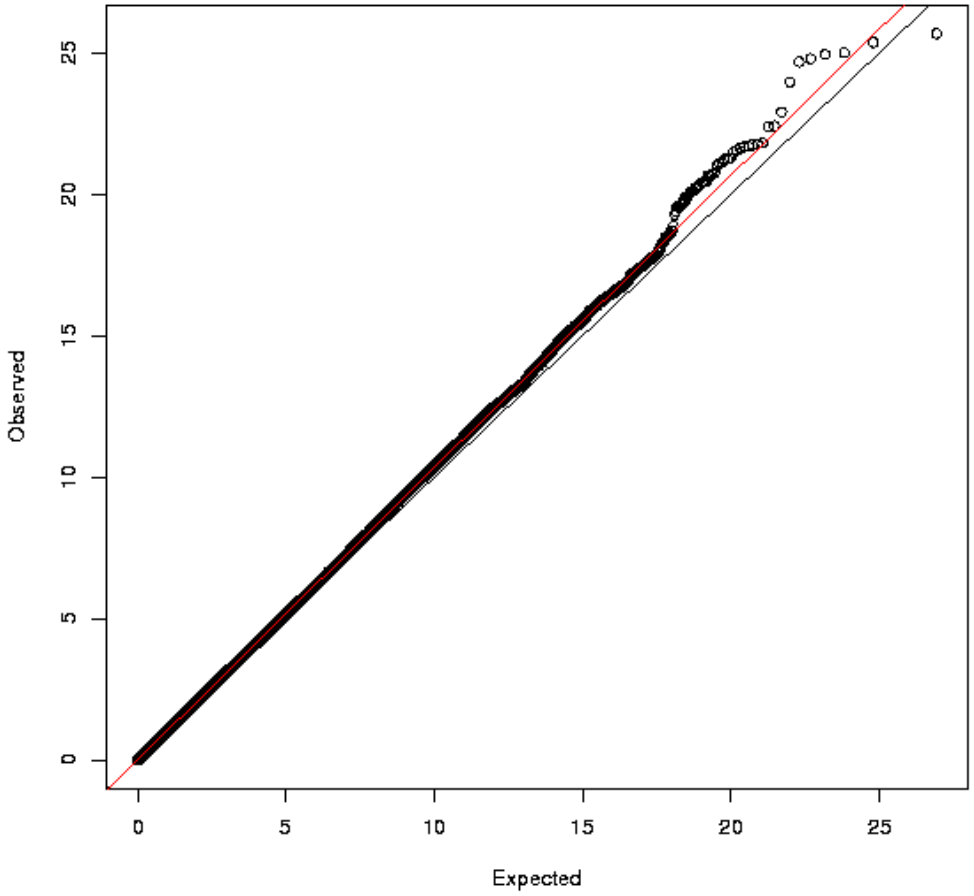
Supplementary Information

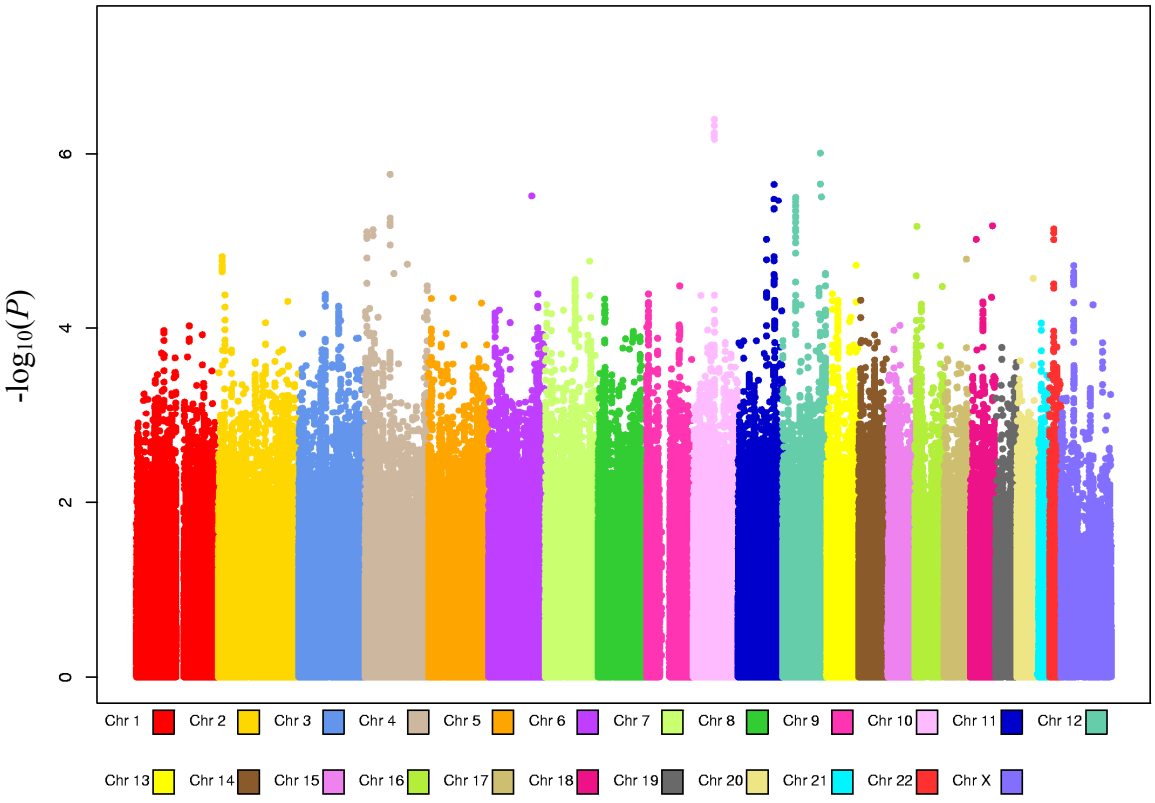
Common variants in ZNF365 are associated with both mammographic density and breast cancer risk

Sara Lindström[1,2], Celine M. Vachon[3], Jingmei Li[4,5], Jajini Varghese[6], Deborah Thompson[6], Ruth Warren[7], Judith Brown[6], Jean Leyland[6], Tina Audley[6], Nicholas J. Wareham[8], Ruth J.F. Loos[8], Andrew D. Paterson[9,10], Darryl Waggott[11], Lisa J. Martin[12], Christopher G. Scott[3], V. Shane Pankratz[3], Susan E. Hankinson[2,13], Aditi Hazra[1,2,13], David J. Hunter[1,2,13], John L. Hopper[14], Melissa C. Southey[15], Stephen J. Chanock[16], Isabel dos Santos Silva[17], JianJun Liu[5], Louise Eriksson[4], Fergus J. Couch[18], Jennifer Stone[14], Carmel Apicella[14], Kamila Czene[4], Peter Kraft[1,2,19], Per Hall[4], Douglas F. Easton[6], Norman F. Boyd[12] and Rulla M. Tamimi[2,13]

Supplementary Figure 1: Quantile-quantile (Q-Q) plot for percent mammographic density. The observed p-values based on meta-analysis of five genome-wide association studies are plotted against the expected distribution of p-values under the null distribution. The genomic inflation factor λ=1.033.



Supplementary Figure 2: Manhattan plot for the meta-analysis of percent mammographic density. The $-\log_{10}(P)$ values are plotted against chromosomal base-pair position. The chromosomes are color-coded.



Supplementary Table 1: Study Characteristics

	Discovery Phase					*Replication Phase*	
Study Acronym	**EPIC-Norfolk**	**NHS**	**SASBAC**	**MBCFS**	**Toronto/Melbourne**	**MCBCS**	**SIBS**
Study	EPIC-NORFOLK	Nurses' Health Study	Singapore and Sweden Breast Cancer Study	Minnesota Breast Cancer Family Study	Toronto/Melbourne	Mayo Clinic Breast Cancer study	Sisters in Breast Screening
Design	Population-based cohort study	Nested breast cancer Case-control study	Retrospective breast cancer case-control study	Family-based breast cancer study	Cross-sectional study of healthy women	Retrospective breast cancer case-control study	Population-based cohort study
Number of Cases/Controls*	0/1,142	806/784	518/740	0/571	0/316	783/907	0/1,145
Mean Age (SD), years	60.1(7.37)	61.2 (6.5)	62.6 (6.2)	57.1 (11.5)	53 (7.4)	60.4 (11.4)	57.5(4.56)
Mean Body Mass Index (SD), kg/m2	28.4 (5.3)	25.8 (4.7)	25.6 (4.0)	27.2 (5.7)	26.5 (3.8)	27.2 (5.5)	27.23(5.2)
% Pre-menopausal women	17.8	0	2.3	30	46	22.7	0.4
% Current HRT user**	24.7	46.1	19.4	28	9	11.2	11
Mammogram view***	MLO	CC	MLO	CC/MLO	CC	CC	MLO
MD measurement software	Cumulus	Cumulus	Cumulus	Cumulus	Cumulus	Cumulus	Cumulus

* Breast cancer cases
** HRT: Hormone Replacement Therapy
*** MLO: Mediolateral-oblique, CC: Cranio-caudal

Supplementary Table 2. Distributions of percent mammographic density according to site and age group for subjects with GWAS data

		40-49 yrs					*50-59 yrs*					*60-70+ yrs*				
		N	Mean (SD)	10th percentile	Median	90th percentile	N	Mean (SD)	10th percentile	Median	90th percentile	N	Mean (SD)	10th percentile	Median	90th percentile
Discovery Phase	**EPIC-Norfolk**	28	29.8 (19.6)	8.5	24.9	58.9	928	21.3 (16.6)	3.79	16.5	46.3	166	12.7 (11.1)	2.45	9	27.63
	NHS controls	16	27.2 (24.0)	1.6	20	55.3	261	27.0 (18.8)	4.4	23.6	53.6	507	22.7 (17.1)	3.7	19	47.4
	NHS cases	24	39.2 (17.6)	19.7	36.4	62.3	265	32.6 (18.9)	8.1	31.4	59.3	517	26.7 (17.7)	5.9	23.5	52.3
	SASBAC controls	1	8.17 (0.0)	8.17	8.17	8.17	250	17.0 (15.0)	2.07	11.9	39.6	489	13.3 (13.4)	1.46	9.03	32.4
	SASBAC cases	0	--	--	--	--	191	23.2 (17.0)	4.98	18.6	49.8	327	14.4 (12.0)	2.1	11.4	30.2
	MBCFS	165*	35.8 (16.6)	14.9	35.1	58.3	166	25.7 (14.4)	7.4	24.7	42.6	240	20.4 (12.9)	5.7	17.8	38.5
	Toronto /Melbourne	538**	36 (22.0)	4.1	35.7	65.7	688	27.2 (19.1)	3.9	24.9	55	385	20.4 (16.3)	2.3	17.6	43.4
Replication Phase	**MCBCS controls**	156***	31.2 (17.0)	10.7	28.8	54	257	22.2 (13.6)	5.5	20.4	38.4	494	18.0 (11.9)	4	16.1	35.3
	MCBCS cases	190****	37.4 (15.5)	18.6	36.2	56.8	229	28.8 (14.3)	11	27.4	48.9	364	22.4 (12.8)	7.5	20.8	39.9
	SIBS	22	37.34 (15.6)	15.28	37.17	56.43	830	20.28 (14.7)	4.28	16.88	42.91	293	16.96 (13.5)	3.07	14.36	31.2

* Including N=32 subjects who are younger than 40
** Including N=34 subjects who are younger than 40
*** Including N=11 subjects who are younger than 40
**** Including N=29 subjects who are younger than 40

Supplementary Table 3. Study-specific details about genotyping, imputation and association analysis

		EPIC-NORFOLK	*NHS*	*SASBAC Cases*	*SASBAC Controls*	*MBCFS*	*Toronto/Melbourne*
GENOTYPING	Genotyping Platform	AffyMetrix 500k	Illumina HumanHap550	Illumina HumanHap240 + HumanHap300	Illumina HumanHap550	Illumina 660W Quad	Illumina HumanHap 1M
	Genotyping Facility	Sanger Institute	NCI Core Genotyping Facility	Genome Institute of Singapore	Genome Institute of Singapore	Mayo Clinic	DeCode Genetics
QUALITY CONTROL	**Sample exclusion due to:**						
	Low call rates	90 (<94%)	59 (<90%)	0 (<90%)	0 (<90%)	2 (<98%)	1 (<99%)
	Heterozygosity	20	0	0	0	0	0
	Ethnicity	13	4	0	0	0	3
	Relatedness	4	0	8	0	0	0
	Others	13	18	0	0	24	0
	SNP exclusion due to:						
	Low call rates	67,192 (<95%)	8,706 (<90%)			2,255 (<95%)	2,400 (<90%)
	Hardy-Weinberg Equilibrium	22,432 (p< 10-5)					
	Low minor allele frequency (MAF)	64,899 (MAF<1%)	18,473 (MAF <1%)			14,455 (MAF=0 %)	172,984 (MAF<1%)
	Others			3,089 (Call rates<90%, MAF<0.03)	34,720 (Call rates<90%, MAF<0.03)		
IMPUTATION	Software	IMPUTE	MACH	IMPUTE	IMPUTE	MACH	PLINK
	HapMap Version (CEU)	HapMap Phase II	HapMap Phase II, Release 21a	HapMap Phase II, Release 21a	HapMap Phase II, Release 22	HapMap Phase II, Release 21a	HapMap Phase II, Release 24
ASSOCIATION	Software	SNPTEST	ProbAbel	SNPTEST	SNPTEST	R – MULTIC	PLINK
	Adjustment factors	Age, number of live births, age at menarche, age at first child birth, Menopausal status, HRT, BMI, chest, hip	Age, BMI, population stratification	Age, BMI, population stratification	Age, BMI, population stratification	Age BMI, menopausal status	Sampling: Age BMI, menopausal status, parity Analysis: Age, BMI, population stratification
	λ GC (Genotyped SNPs)	1.009	0.999	1.045	1.029	1.035	0.999
	λ GC (Imputed SNPs)	1.007	0.997	1.039	1.035	1.029	0.998

Supplementary Table 4. Study-specific results for the top-ranking SNPs at the *ZNF365* locus from meta-analysis

	SNP	***rs10995195***	***rs10995194***	***rs10995190***	***rs10995191***	***rs10995189***	***rs4746419***
	Chr	10	10	10	10	10	10
	Gene	*ZNF365*	*ZNF365*	*ZNF365*	*ZNF365*	*ZNF365*	*ZNF365*
	Position	63958395	63958136	63948688	63948880	63948187	63945267
	Alleles	T/C	G/C	G/A	C/T	G/A	G/C
EPIC-Norfolk*	MAF	0.15	0.14	0.15	0.15	0.15	0.15
	BETA	-0.078	-0.085	-0.077	-0.077	-0.077	0.077
	SE	0.044	0.045	0.044	0.044	0.044	0.044
	P	0.077	0.060	0.078	0.078	0.078	0.078
NHS	MAF	0.16	0.14	0.15	0.15	0.15	0.15
	BETA	-0.301	-0.302	-0.290	-0.289	-0.287	0.288
	SE	0.085	0.087	0.083	0.083	0.083	0.084
	P	0.0004	0.0005	0.0005	0.0005	0.0006	0.0006
SASBAC Cases	MAF	0.14	0.12	0.13	0.13	0.13	0.13
	BETA	-0.264	-0.264	-0.260	-0.260	-0.260	0.255
	SE	0.146	0.148	0.142	0.142	0.142	0.143
	P	0.069	0.073	0.068	0.068	0.068	0.075
SABAC Controls	MAF	0.14	0.13	0.15	0.15	0.15	0.15
	BETA	0.005	0.003	0.007	0.007	0.007	-0.003
	SE	0.115	0.117	0.113	0.113	0.113	0.113
	P	0.971	0.976	0.953	0.953	0.953	0.977
MBCFS	MAF	0.15	0.14	0.14	0.14	0.14	0.15
	BETA	-0.275	-0.259	-0.265	-0.265	-0.265	0.270
	SE	0.110	0.111	0.109	0.109	0.109	0.110
	P	0.013	0.019	0.015	0.015	0.015	0.014
TORONTO**	MAF	0.16	0.15	0.15	0.15	0.15	0.16
	BETA	-0.583	-0.586	-0.560	-0.560	-0.560	0.551
	SE	0.240	0.243	0.234	0.234	0.234	0.235
	P	0.015	0.016	0.017	0.017	0.017	0.019
META-ANALYSIS	Z-Score	5.068	-5.037	-5.001	-4.995	-4.980	4.968
	P ALL	4.02E-07	4.73E-07	5.70E-07	5.87E-07	6.35E-07	6.75E-07
	P HET	0.38	0.44	0.39	0.39	0.40	0.41

* The beta estimates for EPIC-Norfolk are based on a linear regression using log(percent mammographic density) as outcome

** The beta estimates for the TORONTO/MELBOURNE study are based on a logistic regression model based on extreme sampling.

Supplementary Table 5. Association between rs10995190 and breast cancer risk

Cohort	*Analysis**	*No. of individuals*	*OR (95% CI)*	*P*
NHS	Breast Cancer	806 cases and 784 controls	0.88 (0.72-1.07)	0.20
	Breast Cancer adj. % MD	806 cases and 784 controls	0.93 (0.76-1.14)	0.50
SASBAC	Breast Cancer	518 cases and 742 controls	0.91 (0.73-1.14)	0.43
	Breast Cancer adj. % MD	518 cases and 742 controls	0.93 (0.75-1.17)	0.54
MCBCS	Breast Cancer	783 cases and 907 controls	0.78 (0.65-0.95)	0.01
	Breast Cancer adj. % MD	783 cases and 907 controls	0.85 (0.69-1.03)	0.10
Combined	Pooled breast cancer	2,107 cases and 2,433 controls	0.85 (0.76-0.96)	0.008
	Pooled breast cancer adj, %MD	2,107 cases and 2,433 controls	0.90 (0.80-1.01)	0.09

* All analyses were adjusted for age and BMI

Supplementary Table 6: Association between breast cancer loci and percent mammographic density in MODE

SNP	*Gene/Region*	*Non-effect allele/Effect allele*	*Zscore*	*P all (N=4,877)*	*P** Heterogeneity*	*P controls (N=3,553)*	*P cases (N=1,324)*	*Reference*
rs1011970	*CDKN2A,CDKN2B*	G/T	0.65	0.510	0.74	0.578	0.764	Turnbull, Nat. Genet 2010
rs1045485	*CASP8*	G/C	0.66	0.511	0.59	0.538	0.696	Cox, Nat. Genet 2007
rs10941679	*5p12*	G/A	-0.39	0.700	0.53	0.771*	0.920	Stacey, Nat Genet, 2008
rs10995190	*ZNF365*	G/A	-5.00	0.0000006	0.39	0.000033	0.011	Turnbull, Nat. Genet 2010
rs11249433	*1p11.2*	C/T	0.44	0.660	0.02	0.430	0.537*	Thomas, Nat Genet, 2009
rs1219648	*FGFR2*	G/A	-0.12	0.903	0.14	0.787	0.203*	Hunter, Nat. Genet 2007
rs13281615	*8q24.21*	G/A	-0.70	0.482	0.48	0.281	0.570*	Easton, Nature 2007
rs13387042	*2q35*	G/A	-0.49	0.623	0.99	0.815	0.532	Stacey, Nat Genet, 2008
rs1562430	*8q24.21*	C/T	0.02	0.986	0.47	0.877	0.470*	Easton, Nature 2007
rs16886165	*MAP3K1*	G/T	1.34	0.180	0.88	0.271	0.502	Thomas, Nat Genet, 2009
rs2046210	*ESR1*	G/A	2.78	0.005	0.22	0.131	0.003	Zheng, Nat Genet. 2009
rs2380205	*ANKRD16,FBXO18*	C/T	-1.29	0.197	0.43	0.413	0.327	Turnbull, Nat. Genet 2010
rs2981579	*FGFR2*	C/T	-0.26	0.799	0.09	0.946	0.176	Turnbull, Nat. Genet 2010
rs2981582	*FGFR2*	G/A	0.19	0.847	0.25	0.789	0.227*	Easton, Nature 2007
rs3803662	*TNRC9/TOX3*	C/T	1.42	0.157	0.20	0.427	0.171	Easton, Nature 2007
rs3817198	*LSP1*	C/T	-2.09	0.036	0.94	0.105	0.126	Easton, Nature 2007
rs4973768	*3p24*	C/T	1.11	0.267	0.58	0.549	0.277	Ahmed, Nat. Genet 2009
rs614367	*MYEOV,CCND1*	C/T	-0.16	0.876	0.04	0.333*	0.109	Turnbull, Nat. Genet 2010
rs6504950	*17q23.2*	G/A	-0.91	0.361	0.94	0.423	0.508	Ahmed, Nat. Genet 2009
rs704010	*ZMIZ1*	G/A	0.17	0.864	0.14	0.835	0.930	Turnbull, Nat. Genet 2010
rs889312	*MAP3K1*	C/A	-0.07	0.947	0.67	0.686*	0.814	Easton, Nature 2007
rs981782	*5p12*	C/A	-0.13	0.895	0.27	0.883	0.613	Easton, Nature 2007
rs999737	*RAD51L1*	C/T	-1.13	0.259	0.76	0.528	0.374	Thomas, Nat Genet, 2009

* The beta estimate in the stratified analysis is not in the same direction as the overall analysis.
** Heterogeneity between studies

Supplementary Methods

Description of study populations

NHS

The Nurses' Health Study (NHS) was initiated in 1976, when 121,700 US registered nurses aged 30 to 55 returned an initial questionnaire [1]. During 1989 and 1990, blood samples were collected from 32,826 women [2]. As part of the Cancer Genetic Markers of Susceptibility Project (CGEMS) breast cancer scan, 1,145 breast cancer cases and 1,142 controls were genotyped with the Illumina HumanHap500 [3]. For 1,590 of these women - of which 806 were breast cancer cases and 784 were controls - we also had mammographic density measurements. Written informed consent was obtained from all participants. This study was approved by the Committee on the Use of Human Subjects in Research at Brigham and Women's Hospital.

EPIC-Norfolk

The EPIC-Norfolk study participants were initially genotyped for studying body mass [4,5]. Data for EPIC-Norfolk were gathered from two subcohorts of the EPIC-Norfolk study [6]. The first subcohort [n=2,566] included participants randomly selected from the EPIC Norfolk population based cohort of 25,663 men and women of European descent, aged 39-79 years, recruited in Norfolk, UK between 1993-97. The other group is the obese case series set, also derived from the EPIC-Norfolk cohort, consisting of 1,685 individuals with obesity (BMI $\geq$ 30kg/m^2). 1,284 of these cases were non-overlapping with the first sub-cohort. Trained nurses collected blood sample and anthropometric data at the health examination. A Health and Lifestyle questionnaire was completed before the health check. The study protocol was approved by The Norfolk and Norwich Hospital Ethics Committee, and written informed consent was obtained from all participants. Genotyping was done at the Affymetrix services laboratory using the Affymetrix GeneChip Human Mapping 500K array set (Santa Clara, CA, USA).

SASBAC

The Singapore and Sweden Breast Cancer Study (SASBAC) is a population-based case-control study of postmenopausal breast cancer in women born in Sweden aged 50 to 74 years at the time of enrollment, which was between 1 October 1993 and 31 March 1995. Controls were randomly selected from the Swedish Total Population Register and were frequency matched to the expected age distribution of the cases. Details on data collection and subjects have been described previously [7]. The final study group with both mammographic density and genotype data included 571 breast cancer cases and 742 controls. Approval of the study was given by the ethical review board at the Karolinska Institutet (Stockholm, Sweden) and six other ethical review boards in the respective regions in which the subjects were based, and written informed consent was obtained from each participant. Cases and controls were genotyped separately and were therefore treated as two separate populations in this analysis. Breast cancer cases were genotyped using the Illumina HumanHap240 and Illumina HumanHap300 arrays. The controls were genotyped using the Illumina HumanHap550 array.

MBCFS

The families used in this study have been described earlier [8,9]. In total, 426 multigenerational families were ascertained through a breast cancer proband diagnosed from 1944 to 1952 at the University of Minnesota. Probands were consecutive cases, unselected for family history. First- and second-degree blood relatives of the proband and spouses were interviewed between 1990 and 1996; 93% of those contacted provided a telephone interview that included detailed risk factor information. Almost all (99%) women in the 426 families were Caucasian and from Minnesota.

Simulation studies were done to identify the families most informative for linkage analyses. A subset of 90 of the 426 families was selected, and 1,146 family members were invited to provide a blood or buccal sample as a source of DNA; 901 (79%) consented. After the exclusion of 12 individuals due to Mendelian (familial) inconsistencies across markers, the final sample included 89 families, with 889 Caucasian individuals (133 men, 756 women). As part of the parent study, women provided the location of the most recent mammogram and permission to obtain and digitize their mammograms. Mammograms were requested from clinics across the United States, and all were recent mammograms done over the 1990 to 2001 period when national standards were in place for mammography. Among the 737 age-eligible women, we retrieved the mammograms of 658 (89%). Of women with mammograms, 618 (82%) had both craniocaudal and mediolateral oblique views available. Five percent of women had a breast cancer diagnosis during the follow-up period (2000–2002); for these women, mammograms before the diagnosis were used. For this study, a total of 597 women with remaining DNA were genotyped with the Illumina HumanHap 660W Quad array. Written informed consent was obtained from all participants. The protocol was approved by the Mayo Clinic Institutional Review Board.

TORONTO/MELBOURNE

The subjects used in this study were all women of Caucasian origin, without breast cancer, who have participated in previous published studies [10-13] and have provided written informed consent for their samples to be used in genetic research. The selection of subjects according to extreme high and low values for the risk factor of interest maximizes power for detecting genetic variants associated with density. In addition, this strategy has the advantage of creating groups at maximally different risk for breast cancer, of minimizing any misclassification between these groups arising from measurement error, of reducing the number of samples to be analyzed and minimizing the associated costs. Although extreme categories of mammographic density are associated with large differences in body weight, age and menopausal status, selection from the upper and lower 10% of the residuals from regression analysis, after adjustment for these variables, can minimize these differences. This method of selecting subjects preserves large differences in percent density between extreme categories, while minimizing differences in the covariates that are associated with density (unpublished data). Ethics approval was obtained from the University Health Network, Toronto. A total of 316 women were genotyped with the Illumina HumanHap 1M Array at deCode genetics.

Description of mammogram collection and reading

NHS

We collected mammograms as close as possible to the date of blood collection (1989 to 1990). To assess mammographic density, the craniocaudal (CC) views of both breasts were digitized at 261 μm/pixel with a Lumysis 85 laser film scanner, which covers a range of 0 to 4.0 optical density. Mammographic density measurements were conducted using the computer-assisted thresholding software CUMULUS developed at the University of Toronto [14]. We used the average percentage density of both breasts for this analysis. This collection has been described in detail in a previous publication [15].

EPIC-Norfolk

Mammograms for the EPIC women were obtained from Norfolk and Norwich Breast Screening Service records. Mammographic studies were undertaken as part of the UK National Health Service Breast Screening Program; in which women are screened every 3 years by two view (mediolateral and craniocaudal view) mammography between ages 50 and 64 years. The 1,142 women who were successfully genotyped and had mammograms available were used in this analysis (144 of whom were from the obese case series; all analyses were adjusted for BMI). Density readings were made using the computer-assisted program CUMULUS [14].

SASBAC

The process of collecting mammographic density data in this study has been described previously [16]. Film mammograms of the mediolateral oblique view were digitized using an Array 2905HD Laser Film Digitizer (Array Corporation, Tokyo, Japan), which covers a range of 0 to 4.7 optical density. For controls, the breast side was randomized. For cases, the side contralateral to the tumor was used. The density resolution was set at 12-bit spatial resolution. Mammographic density measurements were conducted using the computer-assisted thresholding software CUMULUS [14].

MBCFS

Original mammograms were obtained on 658 women and digitized on a Lumiscan 75 scanner with 12-bit grayscale depth. The pixel size was 0.130 x 0.130 mm2 for both the 18 x 24- and 24 x 30-cm2 films. Mammographic density was estimated for each view using a computer-assisted thresholding program CUMULUS [14,17]. For this study, percent mammographic density from the mediolateral oblique and craniocaudal views were averaged and used as the phenotype.

TORONTO/MELBOURNE

Only original films were used, and all were digitized at a pixel size of 260 μm and a precision of 12 bits using Lumysis 85 digitizers in either Toronto or Melbourne (and sent on compact disk to Toronto). One craniocaudal view of one breast was used for each woman (the side was randomly selected for women in North America, and the right side was used for women in Australia). Mammographic density measurements were conducted by one observer using the computer-assisted thresholding software CUMULUS [14].

Replication

MCBCS

The Mayo Clinic Breast Cancer study (MCBCS) is an Institutional Review Board-approved, on-going clinic-based case-control study initiated in February 2001 at Mayo Clinic, Rochester, MN. The study design has been described previously [18-20]. Cases were women aged 18 years or over with histologically-confirmed primary breast carcinoma recruited within six months of date of diagnosis. Women with a history of cancer (excluding non-melanoma skin cancer) were ineligible. Cases lived in the 6-state region that defines Mayo Clinic's primary service population (Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota). Controls recruited from the outpatient practice of the Divisions of General Internal Medicine and Primary Care Internal Medicine at Mayo Clinic without prior history of cancer (other than non-melanoma skin cancer) were frequency matched on age (5-year age category), race and 6-state region of residence to cases. Written informed consent was obtained from all participants. Case participation was 69% and control participation was 71%.

Mammograms of the contralateral (for cases) or left (for controls) breast performed within five years prior to breast cancer diagnosis (median=22 days) or enrollment date (median=0 days) were obtained and digitized for 940 (50%) cases and 1087 (65%) controls at the time of this analysis. Percent mammographic density was estimated from the craniocaudal mammogram view of the majority of the cases and controls (896 cases and 1033 controls) using CUMULUS [14]; mammograms from 98 cases and controls were excluded due to digital mammogram formats or inability to adequately assess density. Intraclass correlation for estimation of percent density was 0.96.

Genotyping of the rs10995190 SNP was successfully performed on a Sequenom iPLEX platform at the Mayo Clinic. A total of 95 CEPH controls on a Coriell test plate, HAPMAPPT01, were also typed simultaneously to establish genotyping accuracy. Genotyping of the rs10995190 SNP displayed high SNP call rates ($\geq$ 99% for cases, controls and overall), high concordant rate (100%), and no deviation from Hardy-Weinberg equilibrium ($p = 0.95$).

Genotype, percent mammographic density and covariate (age and BMI) data was available for density data were available for 783 cases and 907 controls in the MCBCS and used for the replication study.

SIBS

We used data from the Sisters in Breast Screening study, an ongoing study designed to map genes associated with breast density [21]. Families were identified through the National Health Service breast screening program in the United Kingdom. Eligibility was restricted to families in which two or more female blood relatives (sisters, half sisters, first cousins, or aunt-niece) had had mammographic screening. Families whose member could have screening within 2 y of the recruitment were also included. Written informed consent was obtained from all participants. The study was approved by the local research ethical committee. Study recruitment commenced in October 2002. The current analysis was limited to families whose data including mammographic density measurements were completed by July 2007.

For each participant, all available mammograms were retrieved from the local screening unit and were digitized. The mammograms were scanned using the Array 2905 Laser Film Digitizer and the program DICOM ScanPro Plus Version 1.3E (Array Corp), with 50-µm pixel resolution and 12-bit digitization, and an absorbance of 4.7. For each individual, we aimed to collect the earliest and most recently available mammograms. Mammographic density was measured using the CUMULUS program [14]. Mammograms were analyzed in a random order and the reader was blinded to the sequence of the mammograms and to the visual density evaluation that had been done by another reader.

The *in silico* replication of rs10995190 was based on 1,145 women from 563 families who had been genotyped as part of an ongoing genome-wide association study. Samples were genotyped by Illumina (San Diego) using the Illumina HumanCytoSNP-12 platform. Of 1160 genotyped women, 6 were excluded on the basis of non-European ethnicity, and the twin with the lower call rate was excluded for each of the 9 pairs of monozygotic twins. Rs10995190 was not included on the platform, so the result used in the replication came from data imputed using MACH (HapMap Phase II, release 21, CEU) and analyzed using the Probabel [22] software (r^2=0.99). The analysis was adjusted for age at mammographic examination, BMI, waist hip ratio, HRT use (never, former and current users), age at menopause and menarche. Given the family-based study design, we used the matrix of kinship coefficients between all pairs of individuals (as estimated using 8,236 uncorrelated SNPs) to take into account the non-independence of relatives (the mmscore option in Probabel).

Genotyping, quality control and imputation

The five studies used various genotyping platforms and quality control assessments (Supplementary Table 3). All studies used IBS or IBD measurements to identify and exclude individuals with unexpected relatedness. Since the SASBAC genotyped their breast cancer cases and controls on separate platforms, we imputed and analyzed them separately. This was because SNPs that are imputed accurately from one chip, but poorly from another chip may cause differences in the allele frequency that just reflect allele frequency differences between the haplotype reference panel and the study population [23]. Each study performed imputation to obtain a total of approximately 2.5 million autosomal genotypes using HapMap Phase II CEU samples as reference. To reassure high-quality data for all studies, we excluded all SNPs with a minor allele frequency below 0.01 or an imputation quality score below 0.8 (as defined by the RSQR_HAT value in MACH, the PROPER_INFO in IMPUTE and the information content (INFO) measure in PLINK).

GWAS analysis

Primary association analysis was performed separately for each study. All studies except TORONTO/MELBOURNE used linear regression assuming an additive inheritance model. TORONTO/MELBOURNE selected women in the top and bottom 10% of percent mammographic density adjusted for age, BMI and menopausal status and treated women with high density as “cases” and women with low density as “controls” in a logistic regression model.

For imputed SNPs, the estimated number of effect alleles (ranging from 0 to 2) was used as a covariate. To account for the family structure in MBCFS, we used the

"multic" package as implemented in R. Multic uses a linear mixed effects model, whereby the genetic relatedness among individuals is incorporated into the covariance structure of the random effects [24-26]. The relationships between subjects within the SIBS study were adjusted for using the mmscore option within ProbABEL, based on the estimated genomic kinship matrix [22]. The fixed effect was used for the tests of association and covariate adjustment. The TORONTO/MELBOURNE group used logistic regression where women in the 10% top percentile of percent mammographic density adjusted for age and BMI were treated as "cases" and women in the bottom 10% percentile were treated as "controls". All cross-sectional studies used square-root transformed percent mammographic density as an outcome with the exception of EPIC-Norfolk, where log-transformation was the most appropriate choice. All studies adjusted their analysis for age, BMI and population stratification if necessary. Those studies that had a non-negligible proportion of pre-menopausal women adjusted for menopausal status. Some studies adjusted for additional factors as well, as described in Supplementary Table 3. Population stratification was accounted for by using principal component analysis as described in Price et al [27]. We estimated study-specific inflation factors by calculating the slope from a linear regression between the observed p-values and the expected distribution of p-values under the null hypothesis. For this, we used the "estlambda" function as implemented in the "GenAbel" package for the statistical software R. Study-specific genomic inflation factors ranged between 1.00 - 1.05 for genotyped SNPs and 1.00 - 1.04 for imputed SNPs.

Meta-analysis

Meta-analysis was based on summary statistics from the five different studies including a total of 4,887 Caucasian women. For each SNP, we combined study-specific p-values and direction of association using the METAL software [28]. Weights were proportional to study-specific genomic inflation factors and sample size. To account for the extreme sampling scheme in the TORONTO/MELBOURNE study, we up-weighted this study with a scale factor of 3.51. The scale factor was determined in two ways. We first derived the scale factor theoretically, as 2f(K))/(1-phi(K)), where alpha=(1-phi(K)), is the proportion of the population in each extreme, corresponding to a normal deviate of K, and f is the corresponding normal density. As the TORONTO/MELBOURNE study included women sampled in the top and bottom 10% of their mammographic density distribution, the parameters from the TORONTO/MELBOURNE study were multiplied by a factor of 3.51. We confirmed the scale factor with power calculations by comparing the logistic regression of the extremes with a linear regression on a quantitative trait from the underlying population and calculated the sample size needed to achieve 80% power to detect association for various beta estimates. For a SNP to be considered in the meta-analysis, we required genotyping data from at least 3,000 women. We used Cochran's Q statistic to test for heterogeneity across studies.

Association with breast cancer

We tested for association between rs10995190 and breast cancer risk using those studies that were based on case-control data (NHS, SASBAC and MCBCS). We assumed an additive inheritance model. All analyses were adjusted for age and BMI. We applied two different models in the logistic regression: one without adjustment for percent

mammographic density and one with adjustment for percent mammographic density included as a continuous covariate.

REFERENCES

1. Colditz, G.A. & Hankinson, S.E. The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer* 5, 388-96 (2005).
2. Hankinson, S.E. et al. Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. *J Natl Cancer Inst* 90, 1292-9 (1998).
3. Hunter, D.J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 39, 870-4 (2007).
4. Loos, R.J. et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 40, 768-75 (2008).
5. Willer, C.J. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 41, 25-34 (2009).
6. Day, N. et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer* 80 Suppl 1, 95-103 (1999).
7. Wedren, S. et al. Oestrogen receptor alpha gene haplotype and postmenopausal breast cancer risk: a case control study. *Breast Cancer Res* 6, R437-49 (2004).
8. Vachon, C.M. et al. Strong evidence of a genetic determinant for mammographic density, a major risk factor for breast cancer. *Cancer Res* 67, 8412-8 (2007).
9. Sellers, T.A. et al. Fifty-year follow-up of cancer incidence in a historical cohort of Minnesota breast cancer families. *Cancer Epidemiol Biomarkers Prev* 8, 1051-7 (1999).
10. Boyd, N. et al. Breast-tissue composition and other risk factors for breast cancer in young women: a cross-sectional study. *Lancet Oncol* 10, 569-80 (2009).
11. Boyd, N.F. et al. Heritability of mammographic density, a risk factor for breast cancer. *N Engl J Med* 347, 886-94 (2002).
12. Boyd, N.F. et al. The association of breast mitogens with mammographic densities. *Br J Cancer* 87, 876-82 (2002).
13. Tam, C.Y. et al. Risk factors for breast cancer in postmenopausal Caucasian and Chinese-Canadian women. *Breast Cancer Res* 12, R2 (2010).
14. Byng, J.W. et al. Symmetry of projection in the quantitative analysis of mammographic images. *Eur J Cancer Prev* 5, 319-27 (1996).
15. Tamimi, R.M. et al. Common genetic variation in IGF1, IGFBP-1, and IGFBP-3 in relation to mammographic density: a cross-sectional study. *Breast Cancer Res* 9, R18 (2007).
16. Tamimi, R.M. et al. Birth weight and mammographic density among postmenopausal women in Sweden. *Int J Cancer* 126, 985-91 (2010).
17. Vachon, C.M. et al. Mammographic breast density as a general marker of breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 16, 43-9 (2007).
18. Cox, A. et al. A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* 39, 352-8 (2007).

19. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447, 1087-93 (2007).
20. Kelemen, L.E. et al. Genetic variation in the chromosome 17q23 amplicon and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 18, 1864-8 (2009).
21. Kataoka, M. et al. Genetic models for the familial aggregation of mammographic breast density. *Cancer Epidemiol Biomarkers Prev* 18, 1277-84 (2009).
22. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11, 134 (2010).
23. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11, 499-511 (2010).
24. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62, 1198-211 (1998).
25. Amos, C.I. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 54, 535-43 (1994).
26. de Andrade, M., Atkinson, E.J., Lunde, E., Amos, C.I., Chen, J. Estimating Genetic Components of Variance for Quantitative Traits in Family Studies using the Multic routines. *Technical Report: Series No. 78,* Department of Health Science Research, Mayo Clinic, Rochester, Minnesota (2006).
27. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-9 (2006).
28. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190-1(2010).